DOCUMENT RESUME

ED 074 102 TM 002 465

AUTHOR	Offenberg, Robert M.
TITLE	Evolution of a Bilingual Evaluation.
INSTITUTION	Philadelphia School District, Pa. Office of Research and Evaluation.
PUB DATE	Feb 73
NOTE	17p.; Paper presented at the annual meeting of the American Educational Research Association, February, 1973
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Bilingual Education; Criterion Referenced Tests; Elementary Grades; *Evaluation Methods; Forced Choice Technique; Formative Evaluation; Interviews; Program Descriptions; *Program Evaluation; Research Design; Speeches; Standardized Tests; Summative Evaluation

ABSTRACT

Evaluation of ongoing educational programs must necessarily differ from the basic research design; it must change to meet the changes of the program and its environment. Over the three years of the operation of the Philadelphia "Let's Be Amigos" bilingual program, the kinds of data generated in the program evaluation have evolved in response to the demands of project management, community and intra-school-system relations and the Office of Education. The evaluation of process aspects and product aspects of the program have evolved in opposite directions: (1) evaluation of the pupil performance program outcomes has tended to evolve from informal, criterion-referent approaches to more rigorous experimental designs; and (2) evaluation of processes has tended to evolve from formal methods (observational checklists, forced-choice questionnaires) to less rigorous methods (open-ended questionnaires, interviews, etc.). In the first operational years, assessment of pupils' reading was primarily criterion-referent, involving a word-calling test. The assessment of reading skills was modified after first-year evaluation, first passing through a phase in which an attempt was made to prepare materials-derived, criterion-referent tests to assess more complex skills, and from there to standardized tests. Evaluation of curriculum development has evolved from use of a formal checklist to use of an interview structure with open-ended questions. (KM)

FILMED FROM BEST AVAILABLE COPY

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 074102

TM 002 465

EVOLUTION OF A BILINGUAL EVALUATION

by

Robert M. Offenberg
Research Associate

Read at the 1973 Convention American Education Research Association

OFFICE OF RESEARCH AND EVALUATION
Division of Instructional Research and Development

Edward K. Brown
Director

February 1973

Evolution of the Evaluation of a Bilingual Program

The classical experimental design is a static design. A set of hypotheses is conjectured, and a series of observations or measurements is made which either confirm or deny the validity of these hypotheses. One goal is always to complete the experiment and say "aye" or "nay" at the end. A change in the goals of the experiment should never occur midstream, but only after the data has provided insight into the truth of the original hypotheses.

This model, the basic research model, is perfect for assessing the impact of phenomena like the agricultural experiments and laboratory experiments for which most designs and statistical analyses were developed; phenomena which can be isolated from evolutionary forces like the ones which impinge on educational programs from outside, or which develop within them.

In contrast to a classical experiment is one which is subject to these forces for change. The Philadelphia's Let's Be Amigos bilingual program is one of the first group of projects which required comprehensive evaluation of both project outcomes and project processes under Federal guidelines for program accountability developed for Title VII and VIII of the Elementary and Secondary Education Act. Over the three years that the project has been operational, the kinds of data generated in the program evaluation have evolved in response to the demands of project management, demands of community and intra-school-system relations and the demands of the funding agency, the Office of Education. The pattern of this evolution seems systematic enough to warrant the hypothesis that evaluation of other programs may evolve in similar ways.

The evaluation of process aspects and product aspects of the program appeared to evolve in opposite directions as the project matured: (1) Evaluation of the pupil performance program outcomes has tended to evolve

from informal, criterion-referent approaches to more rigorous experimental designs and (2) evaluation of processes has tended to evolve from formal methods (observational checklists, forced choice questionnaires) to less rigorous methods (open-ended questionnaires, interviews, etc.). In this paper, I will describe the changes occurring in the assessment of one aspect of each of these two types of evaluation. The assessment of reading in the Model School component which serves elementary school English- and Spanish-speaking children is the first. The assessment of the curriculum development process for the project as a whole is the second.

One subject area where there was the clear evolution of product evaluation from soft approaches to rigorous ones was reading. In the first operational years, assessment of pupils' was primarily criterion-referent. A word-calling test was developed in which performance could be assessed directly in terms of the materials presented. Pupils were asked to "call" a sample of the words appearing in the reading series in the order in which they were presented. A sample of the data gathered in the first grade in this way in Figure 1. It shows the percent of pupils at mid-year who could call each of the words in the preprimer. The evaluation plan was to compare the pupil performance with a criterion, and decisions were made on the basis of the outcome. The criterion was that the average word would be recognized by 80% of the pupils. As was reported in an earlier paper (Offenberg, 1971), results obtained were below expected levels for Spanish speakers reading as Spanish reading text, but at expected levels for English speakers reading an English reader. Because the data gathered in this study were directly tied to curriculum materials, they were very useful for program modification. The program personnel found that, because there was a greater-than-necessary level of re-entry of materials, the text used in Spanish was too long for the time allotted to reading instruction in the program.

The finding led to a modification of the use of materials--some stories in which material was virtually all review were omitted, in order to speed up the rate of acquisition of new skills.

The assessment of reading skills was modified after first-year evaluation, first passing through a phase in which an attempt was made to prepare materials-derived, criterion-referent tests to assess more complex skills, and from there to standardized testing (Offenberg 1972 and 1973).

In the first of these phases, instruments were developed to assess more complex skills than the word calling examined in the first year. The instruments attempted to measure the children's abilities to (a) read and understand single words through the matching of words to pictures with pictures (see Figure 2) and (b) to read and understand paragraphs (see Figure 3). To assure that the instrument was a good reflection of skills being taught in the program, they were developed by experienced teachers and supervisors of the project. They prepared items closely related to contents of the texts.

At first it seemed easy to develop "criterion" instruments in the formats shown in the figures. However, when the instruments were used and the results analysed, problems emerged. The tests lacked the qualities which they needed to be useful tools. Examination of the item analyses showed that there was little relationship between anticipated difficulties of items and the number of pupils who correctly completed them. Secondly, the expected relationships between pupils success with the items and total scores was not found. Most items had correlations near zero and a fair number had negative correlations with the total score.

I was considering a major overhaul of the reading test package, when it became apparent that the need for a criterion-referent approach was being supplanted by a greater need for program outcome assessment which had

greater meaning outside the context of project than had the criterion-referent approach. It was now the third operational year of the program, and the third year of funding of bilingual programs under Title VII. I received a phone call from the project director who was in Washington. The call brought home to us for the first time realizations that despite the Office of Education's emphasis on criterion-referent approaches and evaluations designed to be useful for project modification, change to a more traditional experimental design and instrumentation would be needed.

In the phone call the project director said that the Office of Education was preparing for testimony for Congress, and wanted to know if we had any concrete data showing gains or growth brought about by the project. When I provided criterion-referent data showing that, within the limits imposed by the state of our tests' development, the pupils could master most of the skills that project planners claimed they could; the person on the other side of the phone said that Congress did not care about meetings of expectancies--they wanted a number--an amount of gain--which they could use to show that bilingual education was "better" than the education which it replaced. Further discussion with the project director indicated that not only the Federal Government, but also other members of the school community wanted some "hard" data, leading to the decision to overhaul the evaluation of performance in the reading skills area.

The shifting of evaluation to a more classical experimental approach brought with it a major problem. Implied in every classical psychological or educational research design is some baseline or comparison behavior to which the treatment group's behavior can be compared. The laboratory ideal for this comparison is a randomly selected control group. This rarely can take place in an educational setting, especially with

a program as politically potent as bilingual program, which held out the promise greatly improved academic growth for Spanish-speaking children, and which involved the community in its planning. Implementation of a classical design required an answer to the question, "Whose child would be in the control group?"

Fortuitously, the evaluation being conducted in Philadelphia was being conducted by a division of the school system itself, and hence the system could provide some background for a reasonably tight "quasi-experimental" design.

An unpublished study was conducted in 1968, in which all "Spanish Speaking" children in the city of Philadelphia were tested in their mother tongue with standardized tests in reading. The children tested in this group either migrated from a Spanish speaking area, or were children of parents who had migrated. This group provided some "base-line" of pre-program performance against which current performance of Spanish-speaking children could be judged. The regular city-wide testing program provided baselines for English speaking children.

The testing procedure used with the baseline group in 1968 was replicated in the program. The results (See Figure 4) show that performance of Spanish-speaking pupils was substantially greater than that of the baseline, suggesting that the program had enhanced performance of Spanish-speaking pupils. Smaller gains were obtained for English-speaking pupils.

Many of you will recognize the process which I have described as an evolution from a formative evaluation approach to a summative one. The forces which led to the evolution was a change from an approach which provided information for project management, to a need to demonstrate the the value of the program to people outside it at the expense of providing rich product-evaluation data for program modification.

In contrast to the product evaluation, examination of the processes always served project management functions. Its evaluation seems to have moved from formal, structured approaches to more informal ones. The first year's evaluation of program processes was geared to developing clear cut statements of intended program processes--success criteria of program management. In the subsequent years, process evaluation has become geared to answering questions of "how," and "why," in order to gain insight into program operation, and to gather meaningful ways of correcting problems. The early approach was to develop and use high-face-validity questionnaires and checklists which were completed by project personnel expert judges or members of the target groups. An example of this approach was the Curriculum Development checklist. Data from this instrument is shown in Figure 5. The Curriculum Development Checklist embodied the criteria which the project director's staff had set to determine the degree to which developed materials met the internal standards of the project. Once these standards had been specified the coordinator of curriculum development used it to review the work of his subordinates, and guide him in the upgrading of the curriculum preparation process. The data that was produced was to some extent a bi-product of the instruments main function, control of the curriculum preparation process. However, the critical element was not how the instrument's data was used, but that it came from an instrument based on expectancies which were clear enough to be put into questionnaire items which could, for the most part, be answered by a "yes" or "no" comment.

In contrast, Figure 2 shows the instrument used to gain insight into the problems of the curriculum distribution process, it was developed in the thrid operational year. It is a structure for an interview which was conducted

by members of the research staff in order to determine how the materials distribution and tryout could be improved. As can be seen, items in it are open-ended. The instrument served mainly as a guide to assure that the correct topics are discussed, but makes few assumptions as to what the responses will be. In the course of analysing the data response, categories were developed to turn the reactions obtained into countables which could be easily summarized. In contrast to the previous instrument these categories could not be developed until after the answers had been given. Despite the looseness of the data gathering process, the recommendations which came from the interviews were pointed. It was found that (a) teachers needed more specific course outlines and schedules which showed them the specific topics and materials to teach, and (b) the project management needed to clarify their roles and clarify the functions of the curriculum centers of the project.

To summarize, it appears that, unlike the classical experiment, which should be carried to its logical conclusion, evaluation of on going, educational programs must change and adapt to meet the changes of the program and the environment in which they are embedded. The experience of the Let's Be Amigos program suggests that the evaluation of pupil outcomes will become more formal and "experiment-like" and evaluation of process will become more open and informal.

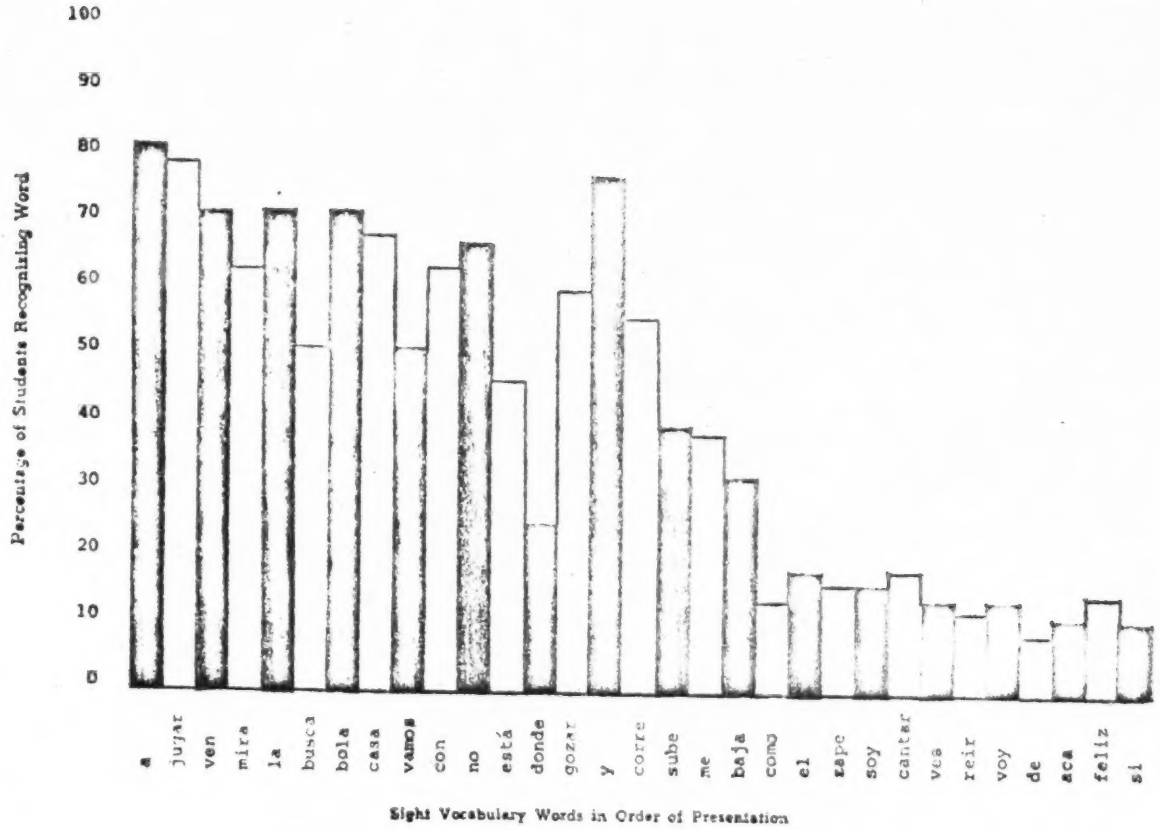
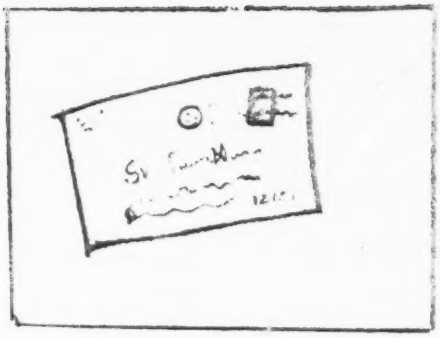


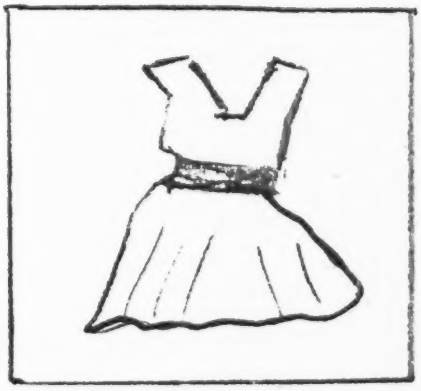
FIG. 1—FEBRUARY READING TEST, LATINO CHILDREN, PRE-PRIMER LEVEL

6.



beso carta vida

7.



música zapato traje

Figure 2. Sample Items from the "Pictures" Criterion Test, Spanish Version (2nd and Third Grade)

12. Kate walked slowly back to her house. In her house, she got some paper and made a sign. Then she took the sign and went back into the street.

Kate took the _____ into the street.

walked house sign

---

13. A big tiger was behind the rocks. A mean-looking tiger. He looked ready to jump.

"Come on". I yelled. "I'm scared! Let's get out of here". We ran the other way.

The boys saw a mean-looking _____.

giraffe getting tiger

Figure 3. Sample Items from the "Paragraph" Criterion Test, English Version (Second-Third Grade).

Comparison of Model School Second- and Third-Grade Pupils' with the Base Line of Latino Pupils in Philadelphia Schools

Grade and Subtest	Base Line (N = 300)			Models A & B (N = 110)		
Second Grade	$\bar{x}$	Percentile[1]	SD	$\bar{x}$	Percentile	SD
Recognition of Words and Letters	43.04	35	15.27	48.70	45	16.20
Word Meaning	8.56	30	4.85	11.61	40	6.59
Comprehension	6.06	36	4.95	7.28	40	7.75
Composite	57.49	32	20.49	67.86	45	26.41
	Base Line (N = 332)			Models A & B (N = 94)		
Third Grade	$\bar{x}$	Percentile	SD	$\bar{x}$	Percentile	SD
Recognition of Words and Letters	49.76	32	13.25	58.39	61	8.48
Word Meaning	9.39	19	8.08	14.25	36	6.51
Comprehension	7.19	21	7.13	10.14	30	6.95
Composite	69.93	27	57.94	82.95	44	19.37

Multivariate Analysis of Variance

Grade Level:	F	df	p<
Multivariate	13.81	4/795	.001
Recognition of Words and Letters	49.87	1/798	.001
Word Meaning	4.41	1/798	.04
Comprehension	9.39	1/798	.002
Composite	18.00	1/798	.001
Program			
Multivariate	20.88	4/795	.001
Recognition of Words and Letters	39.12	1/795	.001
Word Meaning	51.06	1/798	.001
Comprehension	14.93	1/798	.001
Composite	12.29	1/798	.001
Interaction of Grade Level and Program			
Multivariate	1.47	4/795	
Recognition of Words and Letters	1.72	1/798	NS
Word Meaning	2.66	1/798	NS
Comprehension	2.66	1/798	NS
Composite	0.16	1/798	NS

[1]Percentiles are for second- and third-grade rural pupils in Puerto Rico, in the Spring semester.

Figure 4. Sample of the data and analyses used in the "Quasi-Experimental" design phase of Reading Testing.

Summary of supervisor's ratings, on project-developed criteria, of materials completed this year for five curricular units.

Criterion	Number of Units Rated		
	Yes	No	Not Applicable
1. Appropriate for intended grade levels.	5	0	0
2. Appropriate for students' cultural background, interest level, and experiential field.	5	0	0
3. Appropriate for students' previous knowledge in the subject matter or field.	2	0	3
4. Specific objectives clearly stated.	1	4	0
5. Sequential organization and structure.	4	0	1
6. Observable performance outcomes stated.	1	4	0
7. Reasonable variety of learning activities.	5	0	0
8. Evaluation procedures included.	3	2	0
9. Provision for individual rate of learning included.	5	0	0
10. Teacher guide including suggested classroom procedures.	2	3	0
11. Availability of equipment.	1	4	0
12. Aids, materials needed to teach unit specified, and where obtainable.	1	2	2

Figure 5. Sample of the data gathered using the Curriculum Development Checklist.

LET'S BE AMIGOS

Title VII Bilingual Program

Research and Evaluation

Foreign Languages

Structured Interview of Teachers Using
Program Developed Units
Part I Curriculum Distribution

- Identification:

School _____

Teacher's Name _____

Grade Level taught _____

Interviewer _____

Date _____

2. Find out which project developed materials the teacher is using in each subject that he teaches.

Subject	Title	Author
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

3. Find out from whom the teacher got the materials.

Note: (If the teacher does not mention supervisor, the school itself and the curriculum center, ask specifically about them). _____

_____

_____

Note: (If it is not yet clear, find out whether the teacher knows about the Curriculum Development Center at 219 N. Broad, Richard K , and the Materials Center at Potter-Thomas). _____

_____

_____

Figure 6. Sample of the openended type of questionnaire used in evaluation of curriculum distribution.

4. Did the teacher request any materials? If so, what did they ask for, whom did they ask, and did they get them?

5. How can we improve the distribution of materials in general for next year?

6. Next year we would like to examine pupil performance on some of the materials which have been written for use in the project. How can we distribute those materials, and what kind of support can we give to assure that they get a fair trial?

7. Anything else about curriculum materials distribution that we should know?

Figure 6 (Part 2). Sample of the open ended questionnaire used in evaluation of curriculum distribution.

REFERENCES

Offenberg, R. Impact of accountability on the development of a bilingual program. Education, Vol. 93 #1, p. 78.

Offenberg, R. Title VII Project Let's Be Amigos: Evaluation of the First Year. Philadelphia: The School District of Philadelphia. 1971. Reprinted by ERIC, Document ED 046295.

Offenberg, R. Title VII Project Let's Be Amigos: Evaluation of the Second Year. Philadelphia: The School District of Philadelphia, 1972. ERIC in Press.

Offenberg, R. Title VII Project Let's Be Amigos: Evaluation of the Third Year. Philadelphia: The School District of Philadelphia, 1973.